UNIT 6

WAITING LINES/ QUEUING THEORY

OUTLINE

OBJECTIVES

SESSION 6.1:   INTRODUCTION

A body of knowledge about waiting lines, often called queuing theory is an important part of operations and a valuable tool for operations managers. Waiting lines may take the form of cars waiting for repair at a shop, customers waiting at a bank to be served, etc.
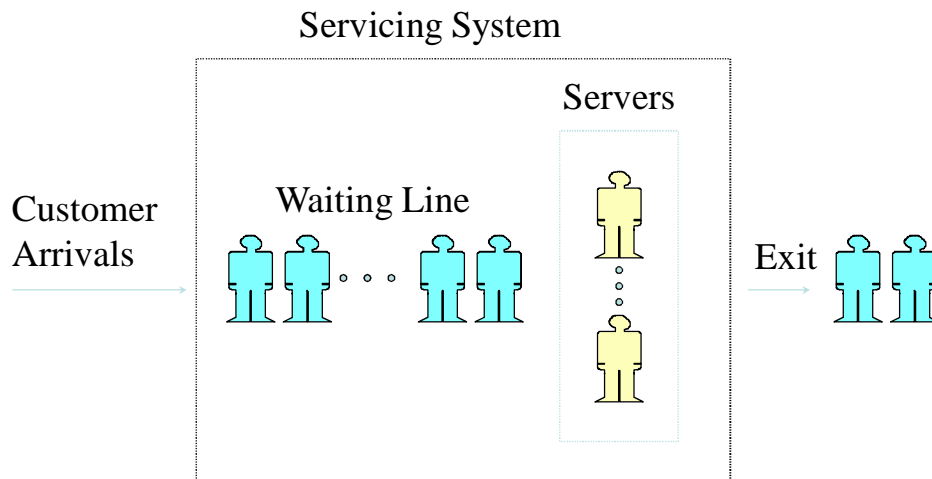
In many retail stores and banks, management has tried to reduce the frustration of customers by somehow increasing the speed of the checkout and cashier lines. Although most grocery stores seem to have retained the multiple line/multiple checkout system, many banks, credit unions, and fast food providers have gone in recent years to a queuing system where customers wait for the next available cashier. The frustrations of "getting in a slow line" are removed because that one slow transaction does not affect the throughput of the remaining customers.

Walmart and McDonald's are other examples of companies which open up additional lines when there are more than about three people in line. In fact, Walmart has roaming clerks now who can total up your purchases and leave you with a number which the cashier enters to complete the financial aspect of your sale. **Disney** is another company where they face thousands of people a day. One method to ameliorate the problem has been to use queuing theory. It has been proved that throughput improves and customer satisfaction increases when queues are used instead of separate lines. Queues are also used extensively in computing---web

servers and print servers are now common. Banks of 800 service phone numbers are a final example I will cite.

Queuing theory leads one directly to the Poisson Distribution, named after the famous French mathematician Simeon Denis Poisson (1781-1840) who first studied it in 1837. He applied it to such morbid results as the probability of death in the Prussian army resulting from the kick of a horse and suicides among women and children. As hinted above, operations research has applied it to model random arrival times.

Components of the Queuing Phenomenon

Servicing System

Servers

Customer Arrivals

Waiting Line

Exit

3

SESSION 6.2:   POISSON DISTRIBUTION

The Poisson distribution is the continuous limit of the discrete binomial distribution. It depends on the following four assumptions:

1. It is possible to divide the time interval of interest into many small subintervals (like an hour into seconds).
2. The probability of an occurrence remains constant thoughout the large time interval (random).

3. The probability of two or more occurrences in a subinterval is small enough to be ignored.
4. Occurrences are independent.

Clearly, bank arrivals might have problems with assumption number four where payday, lunch hour, and car pooling may affect independence. However, the Poisson Distribution finds applicability in a surprisingly large variety of situations. The equation for the Poisson Distribution is: $P(x) = \mu^x \cdot e^{-\mu} \div x!$. The number e in the equation above is the base of the natural logarithms or approximately 2.71828182845904523...

Illustration of the Poisson Distribution

Example 1: On average there are three babies born a day with hairy backs. Find the probability that in one day two babies are born hairy. Find the probability that in one day no babies are born hairy.

Solution: a. $P(2) = 3^2 \cdot e^{-3} \div 2 = .224$ b. $P(0) = 3^0 \cdot e^{-3} = .0498$

Example 2: Suppose a bank knows that on average 60 customers arrive between 10 A.M. and 11 A.M. daily. Thus 1 customer arrives per minute. Find the probability that exactly two customers arrive in a given one-minute time interval between 10 and 11 A.M.

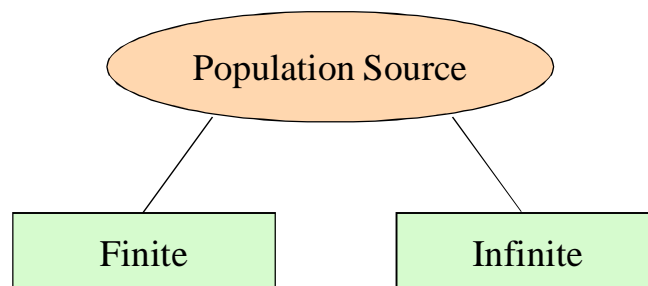**Solution:** Let $\mu = 1$ and $x = 2$. $P(2)=e^{-1}/2!=0.3679\div2=0.1839$.

SESSION 6.3: Characteristics of waiting line systems
The three parts of a waiting-line or queuing system are:
1.   Arrivals or inputs into the system
2.   Queue discipline or the waiting line itself
3.   The service facility

1. **Arrivals or inputs into the system**: The input source that generates arrivals or customers for a service has three major characteristics:

   a. **The size of the source population** The population sizes are considered either unlimited(essentially infinite) or limited(finite). When the number of customers or arrivals on one hand at any moment is just a small portion of all potential arrivals, the arrival population is considered **unlimited or infinite.** It could also be considered as a queue in which virtually unlimited number people or items could request the services. Examples of unlimited populations include cars arriving at highway car toll booth, shoppers arriving at a supermarket, and students arriving to register for classes at a large university. A queue in which there are only a limited number of potential users of the service is termed **finite or limited**

Population Sources

```
                  ┌─────────────────────┐
                  │  Population Source   │
                  └─────────────────────┘
                    /                 \
          ┌──────────────┐      ┌──────────────┐
          │    Finite    │      │   Infinite   │
          └──────────────┘      └──────────────┘
```

4

   b. **The pattern of arrivals at the queue system**: customers arrive at a service facility either according to some known schedule(for example, 1 patient every 15 minutes or 1 student every half hour) or else they arrive randomly. Arrivals are considered random when they are independent of one another and their occurrence cannot be predicted exactly. Frequently, in queuing problems, the number of arrivals per unit of time can be estimated by probability distribution

known as Poisson distribution. For any given arrival time(such as 2 customers per hour or 4 trucks per minute, a discrete Poisson distribution can be established by using the formula:

$$P(x) = \frac{e^{\lambda}\lambda^{x}}{x!} \qquad X = 0, 1, 2, \ldots$$
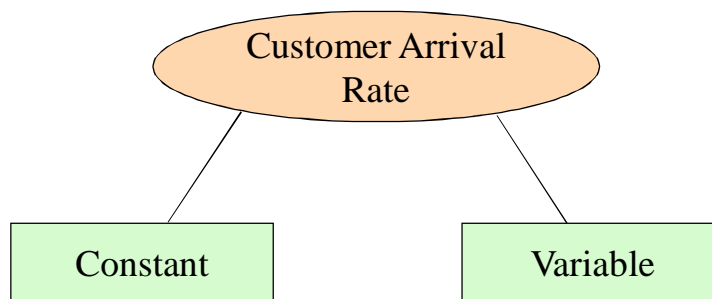
Where:

P(x) = probability of x arrivals

X= number of arrivals per unit of time

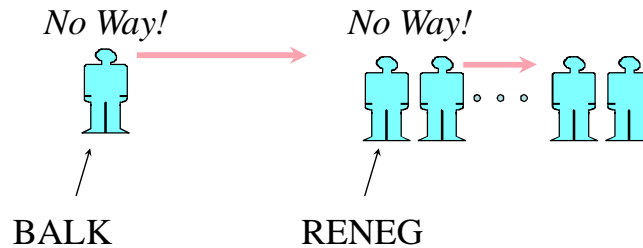λ= average arrival rate

e = 2.7183

Customer Arrival

5

c. **The behaviour of arrivals.** Most queuing models assume that an arriving customer is a patient customer. Patient customers are people or machines that wait in the queue until they are served and do not switch between lines. Unfortunately, life is complicated by the fact that people have been known to **balk or to renege**. Customers who balk refuse to join the waiting line because it is too long to suit their needs or interest. Reneging customers are those who enter the queue but then become impatient and leave without completing their

transaction. Actually, both of these situations just serve to highlight the need for queuing theory and waiting-line analysis.

Degree of Patience

No Way!   *No Way!*

BALK   RENEG

7

2. **Queue discipline or the waiting line itself:** The waiting line itself is the second component of a queuing system. The length of a line can be ether limited or unlimited. A queue is limited when it cannot, either by law or because of physical restrictions, increase to an infinite length. A barbershop, for example, will have only a limited number of waiting chairs. Queuing models are treated in this module under an assumption of unlimited queue length. A queue is unlimited when its size is unrestricted, as in the case of toll booth serving arriving automobiles. A second waiting-line characteristics deals with queue discipline. This refers to the rule by which customers in the line are to receive service. Most systems use a queue discipline known as **first-in, first-out (FIFO-a queuing discipline where the first customers in line receive first service) rule.** In a hospital emergency room or an express checkout line at a supermarket, however, various assigned priorities may pre-empt FIFO. Patients who are critically injured will move ahead in treatment priority over patients with broken fingers

3. **The service facility:** The third part of any queuing system is the service facility. Two basic properties are important.

a. **The configuration of the service system:** Service systems are usually classified in terms of their number of channels(for example, number of servers) and number of phases(for example, number of service stops that must be made). **A single-channel queuing system** with one server, is typified by the drive-in bank with only one open teller. If on the other hand, the bank has several tellers on duty, with each customer waiting in one common line for the first available teller on duty, then we would have **multiple-channel queuing system**. Most banks today are multichannel service systems, as are most large shops, airline ticket counters, and post offices. In a **single-phase system**, the customer receives service from only one station and then exits the system. A fast food restaurant in which the person who takes your order also brings your food and takes your money is a single-phase system. So is a driver's license agency in which the person taking your application also grades your test and collects your license fee. However, say, the restaurant requires you to place order at one station, pay at a second station, and pick up your food at a third station. In this case, it is **multiphase system.** Likewise, if the driver's license agency is large or busy, you will probably have to wait in one line to complete your application(the first service stop), queue again to have your test graded and finally go to a third counter to pay your fee. To help you relate the concepts of channels and phases, see the chart below.

Line Structures

	Single Phase	Multiphase
Single Channel	One-person barber shop	Car wash
Multichannel	Bank tellers' windows	Hospital admissions

6

b. **The pattern of service times:** Service patterns are like arrival patterns in that they may be either constant or random. If service time is constant, it takes the same amount of time to take care of each customer. This is the case in a machine-performed service operation such as an automatic car was. More often, service times are randomly distributed. In many cases, we can assume that random service times are described by negative exponential probability distribution..

SESSION 6.4: Measuring the queue's performance

Queuing models help managers make decisions that balance service costs with waiting line cost. Queuing analysis can obtain many measures of a waiting-line system's performance including the following:

i. Average time that each customer or objects spends in the queue.

ii. Average queue length

iii. Average time that each customer spends in the system(waiting time plus service time)

iv.   Average number of customers in the system

v.   Probability that the service facility will be idle

vi.   Utilization factor for the system

vii.   Probability of a specific number of customers in the system

SESSION 6.5: Queuing Model

All the queuing models that we shall look at have the following characteristics:

i.   Poisson distribution arrivals

ii.   FIFO discipline

iii.   A single-service phase

The different models are summarized in the table below

Model	Name	No. of channels	No. of Phases	Service time pattern	Population size	Queue discipline
A	Single-Channel (M/M/1)	Single	Poisson	Exponential	Unlimited	FIFO
B	Multichannel (M/M/S)	Multi	Poisson	Exponential	Unlimited	FIFO
C	Constant service (M/D/1)	Single	Poisson	Constant	Unlimited	FIFO
D	Limited population	Single	Poisson	Exponential	Limited	FIFO

1. Model A: A single-Channel Queuing model with Poisson arrivals and Exponential service time

The most common case of queuing problem involves the single-channel, or single –server, waiting-line. In this situation, arrivals from single-line to be served by a single station. We assume the following conditions exist in this type of system:

a. Arrivals are served on first-in, first-out(FIFO) basis, and every arrival waits to be served, regardless of the length of the line or queue.

b. Arrivals are independent of preceding arrivals, but the average number of arrivals(arrival rate) does not change.

c. Arrivals re described by a Poisson probability distribution and come from an infinite(or very, very large) population.

d. Service times vary from one customer to the next and are independent of one another, but their average rate is known.

e. Service times occur according to the negative exponential probability distribution.

f. The service rate is faster than the arrival rate.

When these conditions are met, the series of equations shown below can be developed. Example 1 and 2 illustrate how model A(which in technical journals is known as M/M/1 model) may be used.

Queuing formulae for model A: A simple system , also called M/M/1

λ = Mean number of arrivals per time period

μ = Mean number of people or items served per time period

(1).　L_s = Average number of units(customers) in the system $= \dfrac{\lambda}{\mu - \lambda}$

(2).　W_s = Average time a unit(customer) spends in the system

(waiting time plus service time) $= \dfrac{1}{\mu - \lambda}$

(3).　L_q =　Average number of units in the queue $= \dfrac{\lambda^2}{\mu(\mu - \lambda)} =$

(4).　W_q =　Average time a unit spends waiting in the queue $= \dfrac{\lambda}{\mu(\mu - \lambda)}$

(5).　P_0 = Probability of 0 units in the system (that is the service unit is idle) $= 1 - \dfrac{\lambda}{\mu}$

(6).　ρ = Utilization factor for the system $= \dfrac{\lambda}{\mu}$

$P_{n>k}$ =Probability of more than k units in the system, where K is the number of

units in the system $= \left(\dfrac{\lambda}{\mu}\right)^{k+1}$

Example 1

Jones, the mechanic at Golden Muffler Shop, is able to install new mufflers at an average rate of 3 per hour(or about 1 every 20 minutes), according to a negative exponential distribution. Customers seeking this service arrive at the shop on the average of 2 per hour, following a Poisson distribution. They are served on a first-in, first-out basis and come from a very large population of possible buyers.

From the description above, we deduce the operating characteristics of Golden Muffler's queuing system

λ = Mean number of arrivals per time period = 2 cars arriving per hour

μ = Mean number of people or items served per time period = 3 cars serviced per hour

L_s = Average number of units(cars) in the system $= \dfrac{\lambda}{\mu - \lambda} = \dfrac{2}{3-2} = 2$

W_s = Average time a unit(car) spends in the system (waiting time plus service time) $= \dfrac{1}{\mu - \lambda}$

$$= \dfrac{1}{3-2} = 1$$

L_q = Average number of cars in the queue $= \dfrac{\lambda^2}{\mu(\mu - \lambda)} = \dfrac{2^2}{3(3-2)} = \dfrac{4}{3}$

W_q = Average time a car spends waiting in the queue $= \dfrac{\lambda}{\mu(\mu - \lambda)} = \dfrac{2}{3(3-2)} = \dfrac{2}{3}$

P_0 = Probability of 0 units in the system (that is the service unit is idle) $= 1 - \dfrac{\lambda}{\mu}$

ρ = Utilization factor for the system $= \dfrac{\lambda}{\mu} = \dfrac{2}{3}$ =66.7% of time the mechanic is busy

$P_{n>k}$ = Probability of more than k units in the system, where K is the number of units in the system $= \left(\dfrac{\lambda}{\mu} \right)^{k+1}$

k	$P_{n>k} = (2/3)^{k+1}$	Comments
0	0.667	Note that this is equal to 1-P_0 = 1-0.33 = 0.667
1	0.444	

2	0.296	
3	0.198	This implies that there is 19.9% chance that more than 3 cars are in the system
4	0.132	
5	0.088	
6	0.058	
7	0.039	

Model B: Multiple-Channel Model

This is a system in which two or more servers or channels are available to handle arriving customers. We still assume that customers awaiting service from one single line and then proceed to the first available server. Multi-channel, single-phase waiting line are found in many banks today: a common line is formed and the customer at the head of the line proceeds to the first free teller(Appendix).

The multi-channel system presented in Appendix again assumes that arrivals follow a Poisson Probability distribution and that service times are exponentially distributed. Service is first-come, first-served, and all servers are assumed to perform at the same rate. Other assumptions listed for the single-channel model also apply.

Queuing formulae for Model B: A simple system, also called M/M/S

The equations are more complex than those in the single-channel model.

SESSION 6.6: Suggestions for Managing Queues

1. **Determine the acceptable waiting time for your customers**. How long do your customers expect to wait? Set operational objectives based on what is acceptable.

2. **Try to divert your customers' attention when waiting**. Install distractions that entertain and physically involve the customer. Providing music, video or some other form of entertainment may help to distract the customers from the fact they are waiting.

3. **Inform your customers of what to expect**. This is especially important when the waiting time will be longer than normal. Tell them why the waiting time is longer than normal and what you will do to alleviate the queue.

4. **Keep employees not serving the customers out of sight.** Nothing is more frustrating to someone waiting in line than to see employees, who potentially could be serving those in line, working on other activities.

5. **Segment customers by personality types or nature of service.** If a group of customers needs something that can done very quickly, give them a special line so that they do not wait for the slower customers.

6. **Train your servers to be friendly.** Never underestimate the power of a friendly server. Greeting the customer by name, or providing other special attention, can go a long way toward overcoming the negative feeling of a long wait.

7. **Modify customer arrival behavior.** Encourage customers to come during the slack periods. Inform customers of times when they usually would not have to wait. Also tell them when the peak periods are.

8. **Adopt a long-term perspective getting rid of the queues.** Develop plans for alternative ways to serve your customers. Where appropriate, develop plans for automating or speeding up the process in some manner.

SELF ASSESSMENT QUESTIONS

QUESTION 1

There is only 1 copying machine in the student hostel of the polytechnic. Students arrive at the rate of 40 per hour(according to a Poisson distribution). Copying takes an average rate of 40 seconds or 90 per hour(according to an exponential distribution). Compute the following:

a. The percentage of time that the machine is used.

b. The average length of the queue.

c. The average number of students in the system

d. The average time spent waiting in the queue

e. The average time in the system

QUESTION 2

Sam, a vertenary doctor, is running a rabies-vaccination clinic for dogs in the local grade school. Sam can "shoot" a dog every 3 minutes. It is estimated that the dogs will arrive independently and randomly throughout the day at a rate of 1 dog every 6 minutes according to a Poisson distribution. Also assume that Sam's shooting times are exponentially distributed. Compute the following:

  i.   The probability the Sam is idle.
  ii.   The proportion of time that Sam is busy
  iii. The average number of dogs being vaccinated and waiting to be vaccinated
  iv.   The average number of dogs waiting to be vaccinated
  v.   The average time a dog waits before getting vaccinated
  vi.   The average amount of time a dog spends waiting in line and being vaccinated

QUESTION 3

Calls arrive at John DeBruzzi's hotel switchboard at a rate of 2 per minute. The average time to handle each is 20 seconds. There is only 1 switchboard operator at the current time. The Poisson and exponential distributions appear to be relevant in this situation

  a. What is the probability that the operator is busy?
  b. What is the average time that a caller must wait before reaching the operator?
  c. What is the average number of calls waiting to be answered?

QUESTION 4

The Charles Leitle Electronics Company retains a service crew to repair machine breakdowns that occur on average of 3 per day(approximately Poisson in nature). The crew can service an average of 8 machines per day, with time distribution that resembles the exponential distribution

  i.   What is the utilization rate of this service?
  ii.   What is the average downtime for a broken machine?
  iii. How many machines are waiting to be serviced at any given time?

iv.   What is the probability that:

    a.   more than one machine is in the system

    b.   more than two are broken and waiting to be repaired or being serviced?

    c.   More than three are broken and waiting to be repaired or being serviced?

    d.   More that four are broken and waiting to be repaired or being serviced?

QUESTION 5

Gab Moss manages a large Mongomery, Alabama, movie theatre complex called Cinema I, II, III, and IV. Each of the four auditorium plays a different film: the schedule staggers starting times to avoid the large crowds that would occur if all 4 movies started at the same time. The theatre has a single ticket booth and a cashier who can maintain an average service rate of 280 patrons per hour. Service times are assumed to follow an exponential distribution. Arrivals on a normally active day are Poisson-distributed and average 210 per hour.

In order to determine the efficiency of the current ticket operation, Gab wishes to examine several queue –operating characteristics.

i.   Find the average number of moviegoers waiting to purchase a ticket

ii.   What percentage of the time is the cashier busy?

iii.   What is the average time that a customer spends in the system?

iv.   What is the average time spent waiting in line to get the ticket window?

v.   What is the probability that there are more than two people in the system, more than 3, more than 4?

QUESTION 6

The cafeteria line in the university student center is a self-serve facility in which in which students select the items they want and then form a single line to pay the cashier. Students arrive at a rate of about 4 per minute according to a Poisson distributing. The single cashier takes about 12 seconds per customer, following an exponential distribution

    a.   What is the probability that there are more than 2 students in the system? More than 3 students, more than 4?

    b.   What is the probability that the system is empty?

c. How long will the average student have to wait before reaching the cashier?

d. What is the average number in the system?

e. If a second cashier is added and works at the same pace, how will the operating characteristics computed in parts (b), (c), (d), and (e) change?(Assume customers wait in a single line and go to the first available cashier).

QUESTION 7

The administrator at a large hospital emergency room faces the problem of providing treatment for patients who arrive at different rates during the day. There are 4 doctors available to treat patients when needed. If not needed, they can be assigned other responsibilities(such as doing lab tests, reports, X-rays diagnoses) or else rescheduled to work at other hours.

It is important to provide quick and responsive treatment, and the administrator feels that, on the average, patients should not have to sit in the waiting area more than 5 minutes before being see by a doctor. Patients are treated on a first-come, first served basis, and see the first available doctor after waiting in the queue. The arrival pattern for a typical day is as follows:

Time	Arrival rate
9am – 3pm	6 patients/hour
3pm – 8pm	4 patients/hour
8pm - midnight	12 patients/hour

Arrivals follow a Poisson distribution, and treatment times, 12 minutes on the average, follow the exponential pattern.

How many doctors should be on duty during each period to maintain the level of patient care expected?

QUESTION 8

Two technicians, working as a team, monitor a group of 5 computers that run an automated manufacturing facility. It takes an average of 15 minutes (exponentially distributed) to adjust a computer that develops a problem. Computers run for an average of 85 minutes(Poisson-distributed) without requiring adjustments. Determine the following:

a. The average number of computers waiting for adjustments

b. The average number being adjusted

c. The average number of computers not in working order.

QUESTION 9

1. What is a waiting line problem? What are the components in a waiting – line system?

2. Briefly describe three situations in which the FIFO discipline rule is not applicable in queuing analysis

3. Provide 4 examples of four situations in which there is limited , or finite waiting line

4. Most banks have changed from having a line in front of each teller to a situation whereby one line feeds all tellers. Which system is better? Why?